# 基于领域本体的中医语义推理诊断系统

许珠香，江 弋
（厦门大学信息科学与技术学院计算机系 福建 厦门 361005）

【摘 要】：本文基于本体理论建立中医系统知识库,并在此基础上开发智能诊断系统。为了在诊断推理中与用户输入的 症状相匹配,文中采用统计学中的 TFIDF 结合语义思想的方法进行相似度计算排序解决,该系统为中医临床医生提供一个诊疗决策的优良工具。

【关键词】：本体;TFIDF；统计学

## 1、引言

传统中医诊断是通过“望”、“闻”、“问”、“切”等方法获得患者的病情资料,结合以往中医看病经验,采用正确的思维方法进行分析，确定病症的临床表现特点与病情变化规律,为临床预防、治疗提供依据。中医诊断方法的核心力量是“辨证”,传统中医诊断的经验医学模式占据统治地位,辨证主要取决于医生的主观经验和判断,因而不可避免带有个人倾向性,甚至可能做出错误决定。近年来,为克服这种弊端,循证医学模式[2] 应运而生。循证医学是在中医诊治、数据挖掘技术、机器学习和人工智能等多学科交叉结合后产生的,这种运用智能化推理诊断模型更客观，更加不带有个人倾向性,该系统作为中医专家的参考依据,可以有效减少医师诊断错误的机会。

该系统中医数据类别包括中医疾病库、中医穴位库、中医中药库、中医方剂库及中医食疗库等。这些库有个共同的特点就是与中医症状相联系，辨证是“以症为据”,症状体系是中医诊治的基础和依据,经过整理和观察,中医症状具有以下特点:

1)高维[4],中医症状非常繁杂,数据特征较多,噪声较大。在本体库中,出现的症状就有 3 千多个。这些症状未进行二次处理,包含许多同义的症状术语,即使经过二次处理,症状体系的维数也是居高不下的。

2)症状变量具有相关性。在以上众多症状中,存在大量的相关特征、冗余特征,甚至干扰特征。如寒热、寒热不退、寒热往来等症状。

## 2、症状匹配模块(Symptoms Matching Module(SMM))

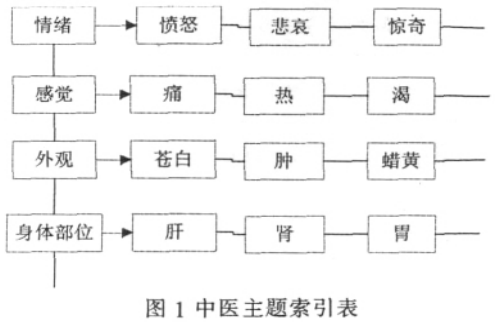


图 1 中医主题索引表

经过观察,大部分症状实体名伴随三种情况出现:一种情况是心理层面(精、气、神),如精神短少、精血亏损等;一种情况是生理层面(五官、脏腑),如舌黄、胸闷、脉细数等；再一种情况是病理层面（排泄及排泄物),如汗泄、便溏、痰黄等[3]。

针对症状实体的这几个特点，对中医症状进行分类,建立中医领域主题词索引表,如图 1 所示。

现需要解决的问题是如何根据用户的输入获取对应的中医症状术语，本文使用基于统计和语义结合的方法。首先计算中医症状术语与用户输入的相似度,采用向量空间模型的TFIDF方法,具体过程如下:

1) 症状词库用向量表示,表示成$(X_1,X_2,...,X_n)$,每个$X_i$表示一个症状成语,如“血枯经闭”,其中n表示词库表中词的总数.每个$X_i$的向量形式为$S=<s_1,s_2,...,s_m>$。$s_i$按下面公式(1)计算:

$$s_i=k_i\times tf_i\times idf_i \quad (1)$$

公式(1)中,$k_i$表示第i个词的权重。中医症候诊断标准中,不同的指标(症状与体征)在辨证中所起的作用大小是不同的,如在表证诊断中,恶寒、发热、头痛、咽痛、脉浮的重要性不同,其中恶寒、脉浮的重要性比其余几个症状大，因此必须考虑各个指标重要性大小不同的问题,即权重。

结合中医症候诊断标准的特点，本文采用确定权重系数的”双百分法” 双百分法是计算某一指标在某一证型中的得分占该指标在各证型中得分总和的百分比”和“认为该证型中可见到该指标的专家数占专家总人数的百分比”,即以这两个百分比的乘积作为该指标在诊断该证型时的权重系数。

$tf_i$表示该词在词库中出现的频率,$idf_i$表示该词在词库中出现的反频率,表示为 $idf_i=\log(N/n)$,N表示词库中症状词总数,n表示包含该词的症状词数。

2) 用户提问的问题,首先分词得到主题词,问题的向量形式为$S'=<s_1',s_2',...,s_m'>$,$s_i'$按下面公式(2)计算:

$$s_i' = k_i' \times tf_i' \times idf_i' \quad (2)$$

公式(2)中，$k_i'$表示第$i$个词的权重，按句首词的权重大于句中、句尾，所以权重逐渐减小，所以权重按位置设为$\frac{1}{i}$，$i$为该词在句中的位置。$tf_i'$表示该词在病人问题库中出现的频率，$idf_i$表示该词在病人问题库中出现的反频率，表示为 $idf_i=log\ (N/n)$，$N$表示库中问题总数，$n$表示包含该词的问题数。

3)设用户提问的问题为S′，词库为S，则计算二者的相似度等同于计算两个向量之间的夹角余弦，如公式(3)所示。

$$Similarity(S,S') = \frac{\sum_{i=1}^{n} s_i \times s_i'}{\sqrt{\sum_{i=1}^{n} s_i^2 \sum_{i=1}^{n} s_i'^2}} = \frac{S \cdot S'}{|S| \times |S'|} \quad (3)$$

基于统计的方法没有考虑到词的语义本身，因此对同义词不能识别。如："大腿肿，怎么办"，在词库中大腿是用"股"表示的，因此定义一张中医同义词库表，基于统计的方法和基于语义的方法，各有所长，把它们的优势结合起来，显著提高系统的性能。方法实现流程如图2所示。

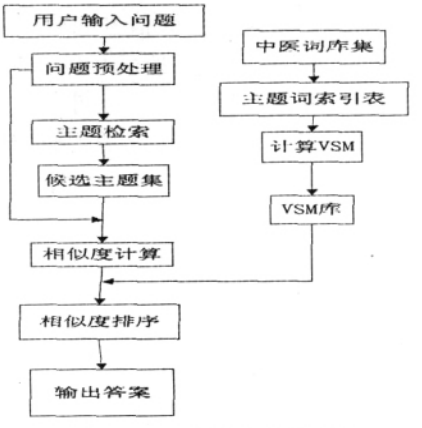


图 2　SMM 模块实现流程图

## 3、Jena 推理实验

**该项工作不是软件工程师能单独完成的，需要和领域专家积极合作，共同建立。以下是该系统的部分规则集[1]：**

rule1:

(?a rdf:type fa:TCM_case),(?b rdf:type fa:TCM_disease),(?a fa:show ?x),(?b fa:hasSys ?x),(?a fa:show ?y),(?b fa:hasSys ?y),(?a fa:show ?z),(?b fa:hasSys ?z),notEqual (?x,?y),notEqual (?x,?z),notEqual (?z,?y) –>(?a fa:has-Disease ?b)

**注:若用户a同时具有各不相同的三种症状x、y、z，疾病b包含这三种症状x、y、z，则a具有疾病b。**

Rule2:

(?a rdf:type fa:TCM_case),(?b rdf:type fa:TCM_meal),(?a fa:show ?x),(?b fa:mealCure ?x)–>(?a fa:canEat ?b)

**注:若用户a同时具有症状x，食物b能治疗症状x，则a适宜吃b。**

Rule3:

(?a rdf:type fa:TCM_case),(?b rdf:type fa:TCM_meal),(?a fa:show ?x),(?b fa:meridianTropism ?x)–>(?a fa:canEat ?b)

**注:若用户a要吃对脏器x有益的食物，食物b归经x，则a适宜吃b。**

Rule4:

(?a rdf:type fa:TCM_case), (?b rdf:type fa:TCM_prescription), (?a fa:show ?x),(?b fa:canCure ?x),(?a fa:show ?y),(?b fa:canCure ?y),notEqual(?x,?y)–>(?a fa:canEat ?b)

**注:若用户a同时具有不相同的两种症状x、y，方剂b能治疗症状x、y，则a适宜吃b。**

**将规则加入Jena推理机[5]中，程序流程图如图(3)所示.**

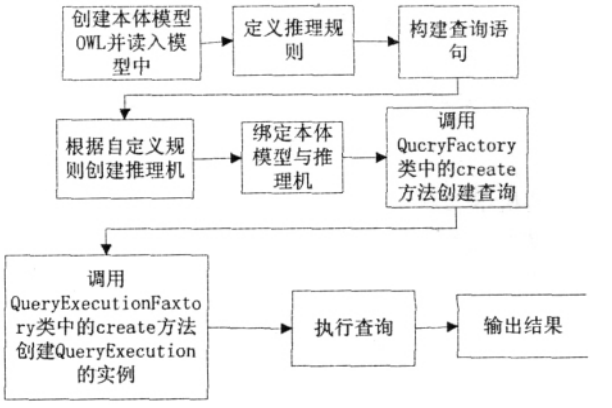


图3 推理程序流程图

**主要程序代码如下所示:**

FileInputStream

```
file=new FileInputStream("E:/webapps/test1/chineseMedicine.owl");
in = new InputStreamReader(file, "UTF-8");
String NS = "http://www.owl -ontologies.com/Ontology1313401328.owl#";
OntModel ontModel =ModelFactory.createOntologyModel (OntModelSpec.OWL_MEM);
ontModel.read(in,null);
String filerule = "E:/webapps/test1/family.rules";
Resource configuration =ontModel.createResource();
configuration.addProperty(ReasonerVocabulary.PROPruleSet, filerule);
Reasoner reasoner = GenericRuleReasonerFactory.theInstance().create(configuration);
InfModel inf = ModelFactory.createInfModel(reasoner, ontModel);
String
prefix = "PREFIX base: <http://www.owl -ontologies.com/Ontology1313401328.owl# > " + "PREFIX rdfs: <http://www.w3.org/2000/01/rdf -schema# > " + "PREFIX rdf: <http://www.w3.org/1999/02/22 -rdf -syntax -ns#>";
String queryString1= " SELECT ?d "+ "WHERE{?z rdf:type base:TCM_symptoms. "
+"?q rdf:type base:TCM_case. "+"?d rdf:type base:TCM_disease. "+"?q base:hasDisease ?d. "+ "}";
queryString2 = " SELECT ?m " + "WHERE {?z rdf:type base:TCM_symptoms. " + "?q rdf:type base:TCM_case. " + "?m rdf:type base:TCM_meal. "+"?q base:canEat ?m. "+ "}";
queryString3 = " SELECT ?f " + "WHERE {?z rdf:type base:TCM_symptoms. "
+"?q rdf:type base:TCM_case. "+"?f rdf:type base:TCM_prescription. "
+"?q base:canEat ?f. "+ "}";
Query query1 = QueryFactory.create(prefix+queryString1);
Query query2 = QueryFactory.create(prefix+queryString2);
```

(下转第 123 页)

为超链接区域，为保证导航整体显示效果，用 CSS 对超链接和块级对象进行一些格式化设置，代码如下：

```
#nav a {
    float:left;
    display:block;
    margin: 0 1px 0 0;
    padding: 4px 8px;
    color:#333;
    text-decoration:none;   //去除超链接的下划线
    border: 1px solid #9b8748;  //设置块级对象的边框样式和颜色
    border-bottom:none;  //不显示选项卡下边框
    background:#f9e9a9;
      }
```


图 2-3　经过 CSS 格式化导航选项卡

4.设置导航选项卡悬停变换[4]

对于选项卡的悬停和选中状态，在<body>中设置 id，名字为 intro，子选择符的功能是基于父对象来选取某个特定对象。通过指定由空格来分开的各个对象，我们可以顺着文档树的顺序缩小目标范围。将悬停和选中两种选项卡变化效果合并在一项 CSS 声明中，代码实现如下：

```
#nav a:hover, body#intro #t-products a{
        color:#333;
        padding-bottom:5px;
        border-color:#727377;
        background:#fff url(img2.jpg) repeat-x top left
        }
```


图 2-4　有悬停变化的导航选项卡

为增强 WEB 导航的设计效果，背景色可以采用图片填充产生立体效果。利用图片处理软件设计两张图片，分别代表选项卡的选中状态和未选中状态。在 CSS 代码中对标签<a>的选中状态和未选中状态进行以下声明：

```
background:颜色值 url(图片文件) repeat-x top/bottom/left/right
```

其中，图片高度任意，宽度设定为 10px，url(图片文件) 表示图片的链接地址，repeat-x 则表示将该图片进行平铺。

至此，我们利用 CSS 和 HTML 对传统的 WEB 导航进行了重构设计，使之对于不同的浏览器、浏览终端和辅助软件有更佳的亲和力。

## 四、结束语

确保 WEB 导航的灵活性是导航设计的关键。本文通过重构普通图片型导航设计方法，使我们摆脱了繁琐的代码束缚。利用 CSS 声明，建立了具有 WEB 可伸缩性导航的概念，为后期站点维护提供了技术保证。

参考文献：

[1]吴建华.浅谈 DIV+CSS 技术[J].福建电脑，2011，(11).
[2]沈茹.CSS 样式表在网页制作中的应用[J].信息与电脑，2011，(01).
[3]吴瑞勇.基于 DIV+CSS 的网页下拉菜单的设计[J].南昌教育学院学报，2011，(01).
[4] 巩恩伟.CSS 在网页开发中的运用技巧 [J]. 中国科技信息，2011，(24).

(上接第 117 页)

```
Query query3 = QueryFactory.create(prefix+queryString3);
QueryExecution qe1 = QueryExecutionFactory.create(query1, inf);
com.hp.hpl.jena.query.ResultSet results1 = qe.execSelect();
QueryExecution qe2 = QueryExecutionFactory.create(query2, inf);
com.hp.hpl.jena.query.ResultSet results2 = qe.execSelect();
QueryExecution qe3 = QueryExecutionFactory.create(query3, inf);
com.hp.hpl.jena.query.ResultSet results3 = qe.execSelect();
```

## 4、小结

基于本体的知识工程是当前的研究热点，本文在这方面作了有益的探索，在以后的工作中，可考虑将病人病史、工作性质、环境气候因素等众多复杂因素进行多信息融合，结合领域专家构造规则集，将所有信息融合成一张错综复杂的网络。

本文作者创新点：结合中医症状特点和用户惯常思维模式，找到通过用户输入内容匹配本体症状库集合的有效途径，以此基础上建立推理诊断模型。

参考文献：

[1] 李新霞.基于本体的中医学脾胃病知识库的构建[D].南京:南京理工大学,2008
[2]王东升,刘亮亮,曹敢,王莉莉,等. 基于领域本体的心血管疾病辅助诊断系统[J].微计算机信息,2008,24(1)
[3]王世昆.中医症状病机实体识别及其关系挖掘研究[D].厦门:厦门大学,2009
[4]王华珍,胡雪琴,等.中医"内生五邪"的智能证型分类[J].计算机工程与应用,2011,47(6)
[5]陈琮.基于Jena的本体检索模型设计与实现[D].武汉:武汉大学,2005
[6]李景.主要本体构建工具比较研究(上)[J].情报理论与实践,2006,29(1):109-111
[7]韩亚洪,刘永革.本体的查询与推理机制研究[J].计算机工程与应用,2005,41(9):82-85
[8]Protégé 新手入门（基础篇）:http://www.docin.com/p-26197625.html
[9] 吴崇胜,陈家旭,胡立胜.Delphi法建立中医证候诊断标准中权重系数确定法新探--双百分法 [J]. 中国中医基础医学杂志,2006,12(4)